Predicting Illness Visits With Neural Networks

David Skrill

12/15/2020

Contents

Introduction

The data considered here are taken from a recently completed study at University of Rochester. The data consists of clinical and molecular measurements and questionaire responses for 166 infants recruited at birth followed for up to three years. The primary goal of this project is to predict whether or not an infant had a respiratory illness at any particular visit, as self-reported by the infant's parent. Modelling was performed using the Least Absolute Shrinkage and Selection Operator (LASSO) and neural networks. The paper preceeds as follows: an introduction to the data and outline of the preprocessing steps taken; presentation and discussion of LASSO results; presentation and discussion of various neural network architectures and results.

Data Preprocessing

The full data consists of the following information:

Description	Observations	Features
Birth medical history	166	21
Birth Demographics	166	5
Family demographics	166	2
oxygen exposure at birth	166	3
Pregnancy medical history	166	47
immune profiling	278	83
Follow up survey 1	635	7
Follow up survey 2	316	32
Target variables and time series info	2339	5
Nasal microbiome	1242	474
Rectal microbiome	1481	490
Throat microbiome	334	128
Virus and bacteria testing	42419	6
vaccination record	2943	4

Subjects are uniquely identified by Alias and a particular visit for an infant is uniquely keyed by the triplet (*Alias, Visit ID, pCGA*), where the Visit ID is a sequential count of previous visits for an infant and pCGA is quantized corrected gestational age. Prior to any processing, the data was split by Alias into training, validation 1, validation 2, and testing data, following a 60%, 15%, 15%, 10% split, respectively.

Selected preprocessing steps are detailed below; for a full account of the preprocessing, see Appendix.

Microbiome and Immune Profiling Data

Each of the microbiome data and the flowcytometry data were filtered to remove all-zero columns and retain distinct Alias, Visit ID, pCGA keys. Missing values were present in the flowcytometry data; these were imputed by Alias using Last Observation Carried Forward (LOCF); any remaining missing values were mean-imputed using column means. Principal component analysis was performed on the resulting tables, producing the following screeplots:

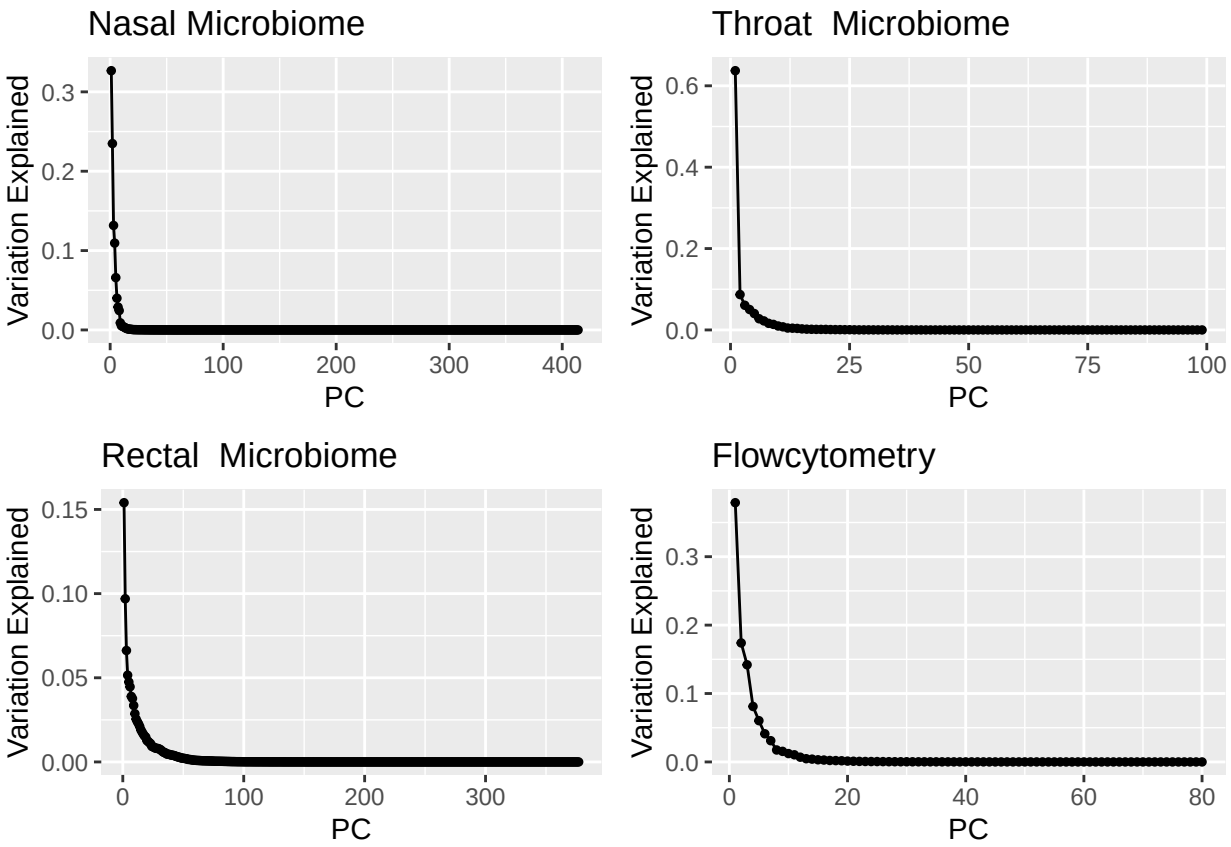


Figure 1: PCA of Microbiome and Flowcytometry Data

The first 50 principal components were retained for the microbiome data; 20 were retained for the flowcytometry data.

Other Proprocessing

One-visit lags were created by Alias for the microbiome, flowcytometry, and virus and bacteria testing data (the longitudinal data). All-missing variables were dropped, and the least-frequent levels for categorical variables were lumped together into an "Other" category, ensuring that "Other" is the smallest category. All data were joined by Alias, Visit ID, and pCGA in cases where such a join was possible; otherwise, the join was done by Alias and pCGA or Alias and Visit ID. Missing values in the resulting table were imputed by Alias, first by LOCB, then LOCF. Any remaining missing values were imputed using the column-wise mean value.

Summary Statistics

% latex table generated in R 4.0.2 by xtable 1.8-4 package % Tue Dec 15 16:36:22 2020

Modeling: LASSO

The first model considered is the logistic LASSO. Penalty parameter values ranged from $\log(-12)$ to $\log(-3)$; the selected parameter, $\log(-4.48)$, maximized the area under the ROC curve produced using predictions on validation 1 data.

Solution Path

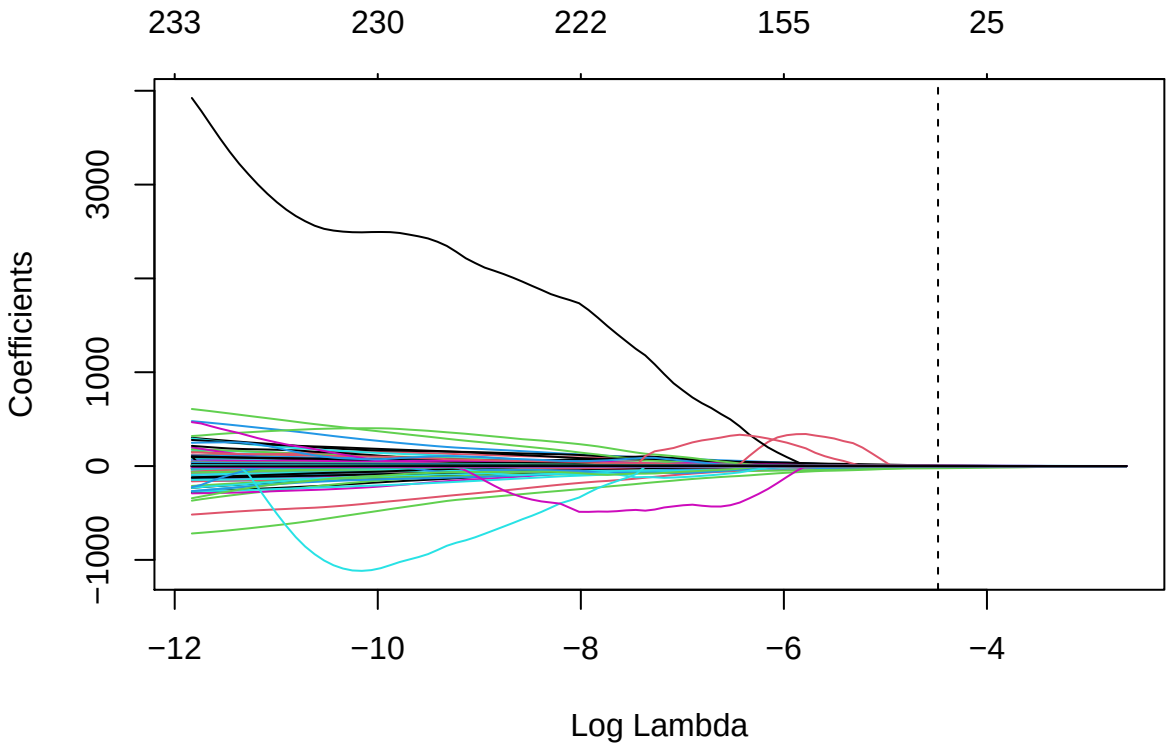


Figure 2: LASSO solution path

It is notable that the selected λ is relatively large, and hence sets many coefficients to 0. At the selected value, there remain 46 coefficients in the model. It is also interesting to note that a single coefficient is extremely large at lower values of λ; it is my suspicion that the training data may be nearly separable along this variable. However, that a higher value of λ was ultimately chosen, thus constraining this coefficient to a more reasonable range suggests that the validation data is probably not seperable in the same way.

Selected Coefficients

The coefficient on PC Nasal lag is unexpectedly large in magnitude; Investigation reveals that one Alias had very low values of this variable during a stretch of illness visits. The selected coefficients confirm what might be expected: that positive or higher bacteria and virus testing results are associated, for the most part, with greater odds of an illness. That so many higher principal components and principal component lags were selected is somewhat surprising, and may suggest that a greater number of principal components be considered. However, given the very small amount of variation explained by any additional principal components, I suspect that doing so would not lead to the detection of generalizable effects.

Evaluation

Loss and AUC Curves

The penalty parameter was chosen to maximize the AUC on validation data. It seems possible that AUC is too coarse a measure to be used for model selection. However, as the following figure shows, the parameter than maximizes the validation AUC is near the minimizer of the log loss on validation data (and testing data as well), lending justification to this choice. As an additional consideration, since the LASSO will also serve as a variable selection method on inputs to neural networks in the next section, it may be desirable to favor a greater number of predictors.

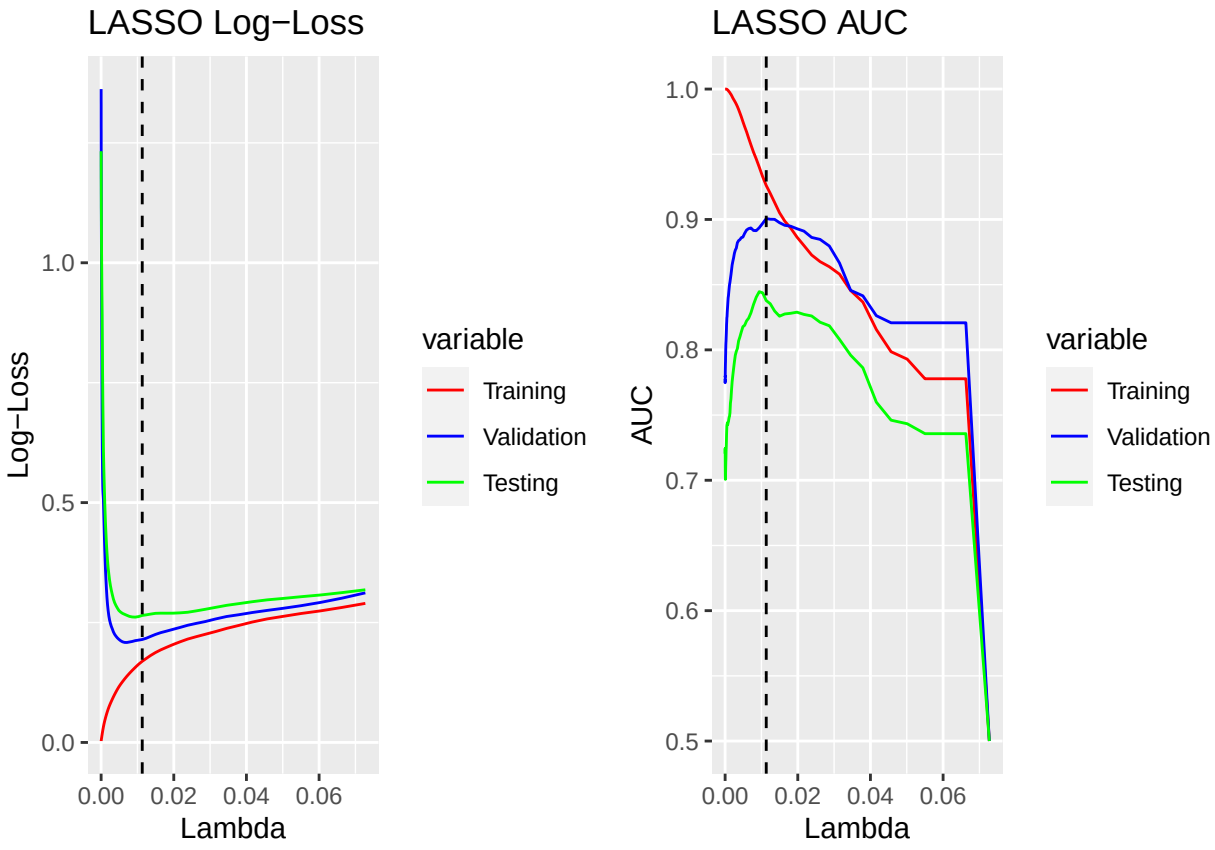


Figure 3: Log-loss and AUC at varying values of lambda. The value that maximizes the validation AUC is represented by the dashed line.

ROC Curves

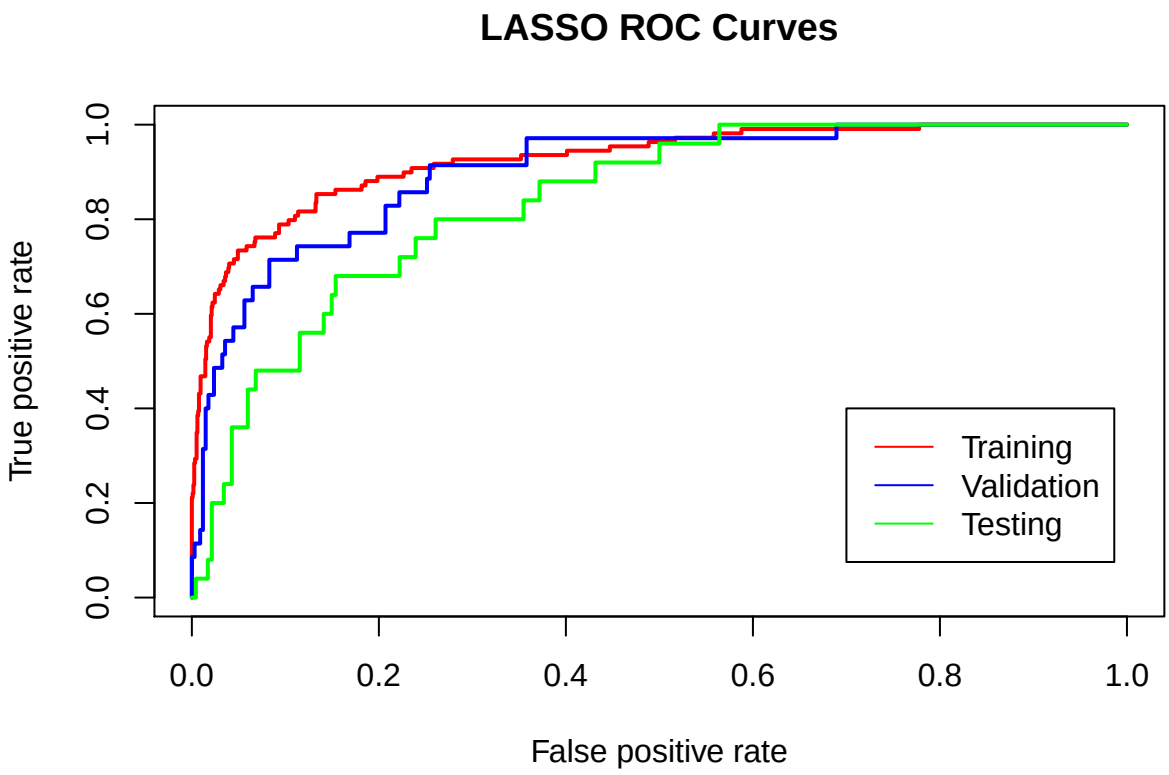


Figure 4: ROC Curves for the logistic LASSO model. The AUCs for training, validation, and testing data are 0.926, 0.901, and 0.838, respectively.

Modeling: Neural Networks

The variables selected by the LASSO were used as inputs to a series of neural networks. The architectures considered were constrained to three hidden layers; 836 architectures were sampled randomly from the following possibilities and assessed on validation 2 data:

- Hidden layer 1 # Nodes: 20, 40, 60
- Hidden layer 2 # Nodes: 20, 40, 60
- Hidden layer 3 # Nodes: 20, 40, 60
- Hidden layer 1 ℓ_2 λ: 0, 0.01
- Hidden layer 2 ℓ_2 λ: 0, 0.01
- Hidden layer 3 ℓ_2 λ: 0, 0.01
- Dropout 1 Rate: 0, 0.2
- Dropout 2 Rate: 0, 0.2
- Dropout 3 Rate: 0, 0.2
- Activation Functions: ReLU, Leaky ReLU, ELU
- SMOTE: Yes, No

In cases where SMOTE was set to yes, the Synthetic Minority Oversampling Technique was used to augment the data. The SMOTE implementation came from the R library DMwR and parameters were set so that for

each instance of an illness visit (the minority class) in the original data, 2 synthetic minority class instances were created; the majority class was sampled at a rate of twice the number of synthetic observations created. The resulting augmented data has a roughly balanced number of illness an non-illness visits.

A chosen activation function was used uniformly throughout, e.g. if ReLU was chosen, it was applied at each hidden layer. The sigmoid activation function was used for the output layer. All networks used a batch size of 128 and were optimized using Adam with an initial learning rate of 0.001. The learning rate was annealed by a rate of 0.2 upon plateau of the validation loss (defined as 10 epochs without improvement). All models were trained for 250 epochs.

Analysis of Network Architectures

A linear model was used to assess the effects of model architecture on validation AUC. All second-order interactions were considered, and the result was processed using AIC-based stepwise selection.

Coefficient	Estimate	SE	t	p
(Intercept)	0.8470	0.0188	45.10	<0.001
nodes1	0.0002	0.0002	0.74	0.458
nodes2	0.0001	0.0002	0.73	0.467
nodes3	-0.0006	0.0003	-2.09	0.037
dropout1	-0.3434	0.0825	-4.16	<0.001
dropout2	-0.0057	0.0420	-0.14	0.892
dropout3	-0.1803	0.0743	-2.43	0.015
lambda1	-0.7515	1.2610	-0.60	0.551
lambda2	0.8979	0.3900	2.30	0.022
lambda3	-3.0341	1.1438	-2.65	0.008
activationLeaky ReLU	-0.1975	0.0158	-12.50	<0.001
activationReLU	-0.0356	0.0157	-2.27	0.023
SMOTETRUE	-0.0455	0.0136	-3.35	<0.001
nodes1:dropout1	0.0023	0.0012	1.94	0.053
nodes1:lambda1	-0.0364	0.0242	-1.51	0.133
nodes2:dropout3	0.0023	0.0012	1.88	0.060
nodes2:SMOTETRUE	-0.0003	0.0002	-1.41	0.158
nodes3:dropout1	0.0019	0.0012	1.59	0.111
nodes3:dropout3	0.0018	0.0012	1.52	0.129
nodes3:lambda3	0.0361	0.0239	1.51	0.132
nodes3:activationLeaky ReLU	-0.0002	0.0003	-0.78	0.438
nodes3:activationReLU	0.0003	0.0003	1.20	0.231
dropout1:dropout2	-0.9040	0.1947	-4.64	<0.001
dropout1:lambda1	-8.8846	3.9062	-2.27	0.023
dropout1:activationLeaky ReLU	0.5460	0.0482	11.32	<0.001
dropout1:activationReLU	-0.2607	0.0483	-5.40	<0.001
dropout1:SMOTETRUE	0.3960	0.0389	10.18	<0.001
dropout2:activationLeaky ReLU	0.1479	0.0477	3.10	0.002
dropout2:activationReLU	-0.1010	0.0475	-2.13	0.034
dropout2:SMOTETRUE	0.1121	0.0390	2.87	0.004
dropout3:lambda3	5.8454	3.8820	1.51	0.133
lambda1:activationLeaky ReLU	1.0632	0.9641	1.10	0.270
lambda1:activationReLU	-1.7215	0.9605	-1.79	0.073
lambda1:SMOTETRUE	4.1444	0.7755	5.34	<0.001
lambda3:SMOTETRUE	1.2675	0.7771	1.63	0.103
activationLeaky ReLU:SMOTETRUE	0.0118	0.0096	1.22	0.222
activationReLU:SMOTETRUE	0.0548	0.0095	5.75	<0.001

It is interesting, if not particularly surprising that the coefficients on the interaction between SMOTE and regularization/dropout parameters are positive: in the augmented data scenario, it is more important to have these in the architecture to combat over-fitting.

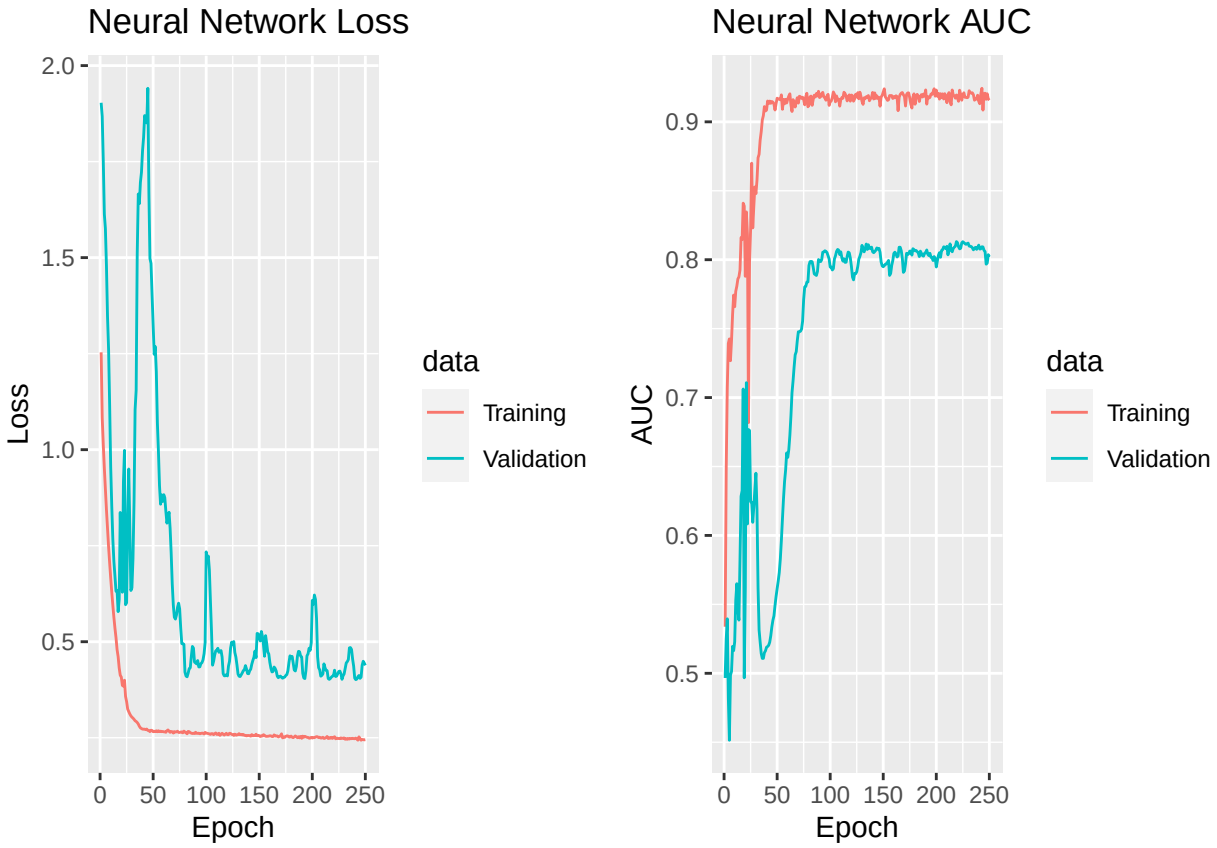
Assessing a Sample Neural Network

Although the above indicates that there are statistically significant effects associated with different model architectures, it is my belief that the top fraction of models are practically interchangable. For example, for the top 100 models (as judged by validation AUC), AUC values ranged from 0.8182 to 0.8477; however, when the top model was retrained, its validation AUC was near the mean validation AUC of the top 100 model group. I believe that much of the variation between these top models results from stochastic optimization, and so little preference should be given to any particular model amoung the top group. As such, the sample model presented here is the model among the top 100 models that is predicted to produce the highest AUC by the model described in the previous section. It has the following architecture:

- Hidden layer 1 # Nodes: 40
- Hidden layer 2 # Nodes: 60
- Hidden layer 3 # Nodes: 20
- Hidden layer 1 ℓ_2 λ: 0
- Hidden layer 2 ℓ_2 λ: 0.01
- Hidden layer 3 ℓ_2 λ: 0
- Dropout 1 Rate: 0
- Dropout 2 Rate: 0
- Dropout 3 Rate: 0
- Activation Function: ReLU
- SMOTE: No

Loss and AUC Curves

As shown in the figure below, substantial and lasting improvement in validation AUC came only after a precipitous drop in training loss and spike in validation loss. This was very common throughout all of the top models. It is interesting that the region of greatest validation AUC improvement comes when training loss is already near its minimum value. There is little improvement after roughly 100 epochs; I suspect that the optimization has found a local minimum, and is prone to doing so due to the annealing of the learning rate.



ROC Curves

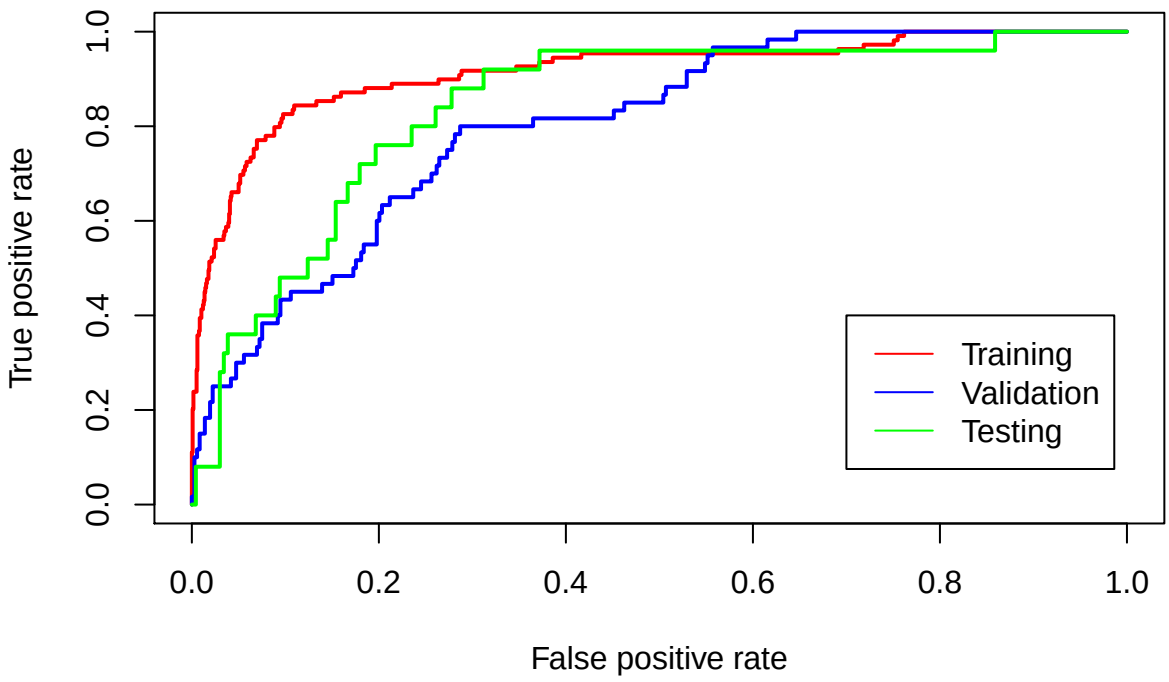


Figure 5: ROC Curves for a sample neural network. The AUCs for training, validation, and testing data are 0.916, 0.802, and 0.843, respectively.

Appendix

Initial preprocessing was performed table-by-table, as described below. These preprocessed tables were then joined, and missing values were imputed using a combination of LOCB, LOCF, and mean-inputation.

Base Too few births occurred outside of the study center to be of use as a predictor; this variable was dropped from the table. Birth Order is NA when the birth was not a multiple birth; NAs were set to 1. All babies that received a stabilization procedure received supplemental oxygen; supplemental oxygen was dropped from the table. For other stabilization procedures, NAs were set to "No", so that the resulting variable is binary ("Yes" or "No").

Preg There were many variables that were all NA; these variables were dropped from the table.

Fup1 There were many NAs in variables `Is baby receiving breast milk exclusively` and `For how many months did baby receive breast milk for more than half of feedings`, and many participants were completely unobserved. Even with imputation, these variables would be problematic; they were dropped from the table.

This table does not contain Visit ID information, posing a potential problem in future joins with tables

keyed by Alias and Visit ID; rows with distinct Alias-pCGA pairs were kept and the remainder discarded. Additionally, rows containing NAs were dropped.

Fup2 An initial screen of variables was performed and variables with fewer than 10 NAs were retained. The variable `How many smokers in home` was set to 0 if the participant reported that smoking is not allowed inside the home at all. Missing values in `Dogs, cats, and other furry animals` and `Kerosene heater, wood burning stove, or fireplace` were set to "No" so that the resulting variables are binary. Again, rows with NAs were then dropped, and and only a distinct Alias-pCGA pairs were retained.

Microbiome Data Each microbiome table was preprocessed in the following steps: all-zero columns were removed, and rows were filtered to retain distinct Alias-Visit ID-pCGA keys; PCA was performed on the subset of columns with names containing "k___Bacteria"; the first 50 principal components were selected.

A set of lag variables was created by Alias; the values at the first record for each Alias were set to zero.

Flowcytometry All-zero columns were discarded, as were rows with no Visit ID information. Again, rows were filtered to retain distinct Alias-Visit ID-pCGA keys. Missing values were imputed using LOCF by Alias; any remaining missing values were imputed using mean-imputation over the full table. PCA was performed on the subset of columns with names containing "Meta.Cluster"; the first 20 principal components were selected. A set of lag variables was created by Alias; the values at the first record for each Alias were set to zero.

TLDA A set of lag variables was created by Alias; the values at the first record for each Alias were set to zero.

Vaccines An additional variable, `vaccinated` was created and set to 1 for all participants in the table. Missing values were imputed using LOCF by Alias, and any remaining NAs were set to 0. The result was filtered to retain distinct Alias-pCGA keys.

Joins All tables were joined by Alias, Visit ID, and pCGA in cases where such a join was possible; otherwise, the join was done by Alias and pCGA or Alias and Visit ID. Missing values in the resulting table were imputed by Alias, first by LOCB, then LOCF. Any remaining missing values were imputed using the column-wise mean value.

Additional Preprocessing All character variables except Alias, Visit ID, and pCGA were set to factors levels were combined using `fct_lump_lowfreq()`; variables that are binary (both those that were originally binary and those that are effectively binary after the combination of levels) were encoded as 0-1; variable names were cleaned using janitor::clean_names().

	nobs	uniq	mean	SD	min	25%	50%	75%	max
Birth Weight (g)	1288	82	2066.23	1135.32	561.00	1120.00	1710.00	3014.00	5049.00
TPiece Resuscitator	1288	2	0.37	0.48	0.00	0.00	0.00	1.00	1.00
Cohort 26-27 weeks	1288	2	0.10	0.31	0.00	0.00	0.00	0.00	1.00
Mother's Edu., Other	1288	2	0.21	0.41	0.00	0.00	0.00	0.00	1.00
Placental Pathology Obtained	1288	2	0.22	0.42	0.00	0.00	0.00	0.00	1.00
Visit ID	1288	22	8.30	4.69	1.00	4.00	8.00	12.00	22.00
PC1 Flow	1288	158	0.05	0.13	-0.39	-0.01	0.05	0.12	0.34
PC3 Flow	1288	158	0.01	0.08	-0.22	-0.03	0.01	0.03	0.48
PC4 Flow	1288	158	0.00	0.06	-0.22	-0.01	0.00	0.03	0.27
PC7 Flow	1288	158	-0.01	0.03	-0.10	-0.02	-0.01	0.01	0.43
PC14 Flow	1288	158	0.00	0.01	-0.08	-0.00	0.00	0.01	0.05
PC18 Flow	1288	158	0.00	0.01	-0.04	-0.00	0.00	0.00	0.03
PC1 Flow lag	1288	97	-0.05	0.10	-0.39	-0.07	-0.05	0.00	0.28
PC5 Flow lag	1288	97	0.00	0.06	-0.17	-0.01	0.00	0.00	0.30
How many smokers in home?	1288	4	0.21	0.55	0.00	0.00	0.00	0.00	3.00
PC3 Nasal	1288	673	0.02	0.22	-0.42	-0.11	-0.00	0.08	0.78
PC16 Nasal	1288	673	-0.00	0.02	-0.08	-0.00	0.00	0.00	0.42
PC3 Nasal lag	1288	589	0.01	0.19	-0.42	-0.06	0.00	0.03	0.78
PC5 Nasal lag	1288	589	0.00	0.13	-0.86	0.00	0.00	0.04	0.39
PC15 Nasal lag	1288	589	-0.00	0.02	-0.10	-0.00	0.00	0.00	0.34
PC31 Nasal lag	1288	589	0.00	0.01	-0.10	0.00	0.00	0.00	0.04
PC39 Nasal log	1288	589	0.00	0.00	-0.03	-0.00	0.00	0.00	0.05
PC45 Nasal lag	1288	589	0.00	0.00	-0.03	-0.00	0.00	0.00	0.09
PC49 Nasal lag	1288	589	-0.00	0.00	-0.02	-0.00	-0.00	0.00	0.03
PC1 Rectal	1288	816	-0.01	0.15	-0.81	-0.05	0.01	0.09	0.23
PC41 Rectal	1288	816	-0.00	0.02	-0.20	-0.01	0.00	0.01	0.14
PC43 Rectal	1288	816	-0.00	0.02	-0.11	-0.01	-0.00	0.01	0.13
PC44 Rectal	1288	816	0.00	0.02	-0.13	-0.01	0.00	0.01	0.13
PC50 Rectal	1288	816	-0.00	0.02	-0.12	-0.01	0.00	0.01	0.10
PC7 Rectal lag	1288	729	-0.00	0.06	-0.46	-0.02	0.00	0.03	0.25
PIV2	1288	2	0.00	0.04	0.00	0.00	0.00	0.00	1.00
RSV	1288	2	0.02	0.13	0.00	0.00	0.00	0.00	1.00
Rhino	1288	2	0.21	0.41	0.00	0.00	0.00	0.00	1.00
MHom	1288	2	0.01	0.10	0.00	0.00	0.00	0.00	1.00
BPert	1288	2	0.00	0.03	0.00	0.00	0.00	0.00	1.00
PIV3	1288	2	0.01	0.11	0.00	0.00	0.00	0.00	1.00
HMPV	1288	2	0.00	0.06	0.00	0.00	0.00	0.00	1.00
Corona 3	1288	2	0.01	0.12	0.00	0.00	0.00	0.00	1.00
MPneu	1288	2	0.00	0.03	0.00	0.00	0.00	0.00	1.00
Parecho	1288	2	0.01	0.08	0.00	0.00	0.00	0.00	1.00
Corona 2	1288	2	0.02	0.13	0.00	0.00	0.00	0.00	1.00
RSV lag	1288	2	0.02	0.13	0.00	0.00	0.00	0.00	1.00
MHom lag	1288	2	0.01	0.10	0.00	0.00	0.00	0.00	1.00
PIC 3 lag	1288	2	0.01	0.11	0.00	0.00	0.00	0.00	1.00
MPneu lag	1288	2	0.00	0.03	0.00	0.00	0.00	0.00	1.00
SPneu lag	1288	2	0.18	0.39	0.00	0.00	0.00	0.00	1.00

Table 2: Selected Coefficients

Coefficient	Estimate
Intercept	-4.3691312
Birth Weight (g)	-0.0000266
TPiece Resuscitator	0.1040783
Cohort 26-27 weeks	0.1296478
Mother's Edu., Other	0.1703382
Placental Pathology Obtained	-0.0185050
Visit ID	0.1420775
PC1 Flow	-0.7565810
PC3 Flow	-0.5249234
PC4 Flow	-3.2640313
PC7 Flow	0.0032479
PC14 Flow	-4.7801318
PC18 Flow	-3.6594495
PC1 Flow lag	2.6290042
PC5 Flow lag	-1.1582453
How many smokers in home?	0.2054745
PC3 Nasal	-0.4436922
PC16 Nasal	4.6134775
PC3 Nasal lag	-0.2034761
PC5 Nasal lag	-0.5321147
PC15 Nasal lag	3.2737969
PC31 Nasal lag	-23.4794439
PC39 Nasal log	2.6028722

Coefficient	Estimate
PC45 Nasal lag	5.1158045
PC49 Nasal lag	-2.7069932
PC1 Rectal	0.0653280
PC41 Rectal	2.2736912
PC43 Rectal	-4.1752421
PC44 Rectal	1.5773573
PC50 Rectal	0.6866258
PC7 Rectal lag	-0.5596358
PIV2	0.6325372
RSV	2.7673806
Rhino	0.8946950
MHom	1.0773441
BPert	0.7943976
PIV3	2.4769241
HMPV	1.7518077
Corona 3	2.1630711
MPneu	2.5927236
Parecho	0.9148245
Corona 2	0.3952674
RSV lag	-1.6047078
MHom lag	0.0623605
PIC 3 lag	-0.5996385
MPneu lag	2.4785912
SPneu lag	0.2968816